

LIBRARY OF THE
UNIVERSITY OF ILLINOIS
AT URBANA-CHAMPAIGN

510.84

IL6r

no. 391-396

cop. 2



The person charging this material is responsible for its return to the library from which it was withdrawn on or before the **Latest Date** stamped below.

Theft, mutilation, and underlining of books are reasons for disciplinary action and may result in dismissal from the University.

UNIVERSITY OF ILLINOIS LIBRARY AT URBANA-CHAMPAIGN

OCT 7 1975
SEP 15 REC'D



Digitized by the Internet Archive
in 2013

<http://archive.org/details/implicitmethodfo392brac>

10.84
6r
392
p 2

AN IMPLICIT METHOD FOR THE SOLUTION
OF ELLIPTIC PARTIAL DIFFERENTIAL EQUATIONS

by

Amnon Bracha

April 15, 1970



THE LIBRARY OF THE

NOV 9 1972

UNIVERSITY OF ILLINOIS
AT URBANA-CHAMPAIGN

DEPARTMENT OF COMPUTER SCIENCE
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN · URBANA, ILLINOIS

Report No. 392

AN IMPLICIT METHOD FOR THE SOLUTION
OF ELLIPTIC PARTIAL DIFFERENTIAL EQUATIONS^{*}

by

Amnon Bracha

April 15, 1970

Department of Computer Science
University of Illinois
Urbana, Illinois 61801

* This work was supported in part by the National Science Foundation under Grant No. US NSF-GJ-328 and was submitted in partial fulfillment for the degree of Master of Science in Computer Science, June, 1970.

ACKNOWLEDGMENT

I would like to express my appreciation to Professor P. E. Saylor, who suggested the thesis topic and supervised its development.

This thesis is dedicated to the memory of my mother.

TABLE OF CONTENTS

	Page
1. INTRODUCTION.	1
2. PRELIMINARIES	3
3. THE FACTORIZATION	7
4. ANALYSIS AND CONVERGENCE CONDITIONS	14
5. COMPUTATIONAL NOTES	26
LIST OF REFERENCES.	28

1. INTRODUCTION

Let $Ax = b$ be a linear system of algebraic equation resulting from the discretization of an elliptic boundary value problem. A direct method for computing the solution typically is equivalent to factoring A as the product $A = LU$ of a lower triangular matrix L by an upper triangular matrix U . Such methods are computationally inefficient because the factors L and U are not sparse matrices. For this reason, an iterative procedure is generally used rather than a direct method. Iterative procedures have recently been proposed by several people, notably H. L. Stone [5], and Dupont, Kendall and Rachford [2], based on the idea of finding a matrix $A + B$ that can be factored as the product of sparse matrices L and U ,

$$A + B = LU.$$

The efficient solution of $(A + B)u = b$ is then used in the iterative scheme defined by

$$(A + B)u_{n+1} = (A + B)u_n - \tau(Au_n - b),$$

where τ is a parameter. Stone refers to such procedures as strongly implicit (SI) procedures.

The subject of this thesis is an SI procedure due to Stone [4]. The factorization in the procedure depends on a parameter α , and we indicate this dependence below with a subscript. Stone designed the factorization to make $L_\alpha U_\alpha = A + B_\alpha$ symmetric. The main result of this thesis is that $A + B_\alpha$ is defined and positive definite for α satisfying $0 \leq \alpha \leq 1$, when A is derived from a discretization of the Dirichlet problem. We also show that,

for A and $A + B_{\alpha}$ positive definite, the procedure

$$(A + B_{\alpha})u_{n+1} = (A + B_{\alpha})u_n - \tau(Au_n - b)$$

converges for values of τ satisfying $0 < \tau < \frac{2}{\|A\|_2}$.

2. PRELIMINARIES

Consider the Dirichlet problem

$$(2.1) \quad -\frac{\partial}{\partial x_1} (a_1(x) \frac{\partial u}{\partial x_1}) - \frac{\partial}{\partial x_2} (a_2(x) \frac{\partial u}{\partial x_2}) = Q(x), \quad x \in D$$

$$u(x) = 0, \quad x \in \partial D$$

where $x = (x_1, x_2)$, D is the interior of a compact region, and where the differential operator in (2.1) is elliptic, that is,

$$a_1(x) \xi_1^2 + a_2(x) \xi_2^2 > 0$$

for

$$(\xi_1, \xi_2) \neq (0, 0) \quad \text{and} \quad x \in D.$$

Cover the region D with a rectangular grid system, replace the differential operator in (2.1) with a difference operator, and replace u with a vector whose components correspond to the grid points of the system. The result is a set of simultaneous linear equations with a one-to-one correspondence between the gridpoints and the equations. We shall study a method for the iterative solution of this set of equations.

In order to make the problem more specific, let D be the unit square,

$$D = \{(x_1, x_2) : 0 < x_1, x_2 < 1\}.$$

Let $h = \frac{1}{n+1}$, n a positive integer. Let D_h be the grid system over D , with uniform spacing h , defined by

$$D_h = \{(jh, kh) : 1 \leq j, k \leq n\}.$$

Denote the grid point (jh, kh) by (j, k) . The boundary, ∂D_h , of D_h is defined by

$$\partial D_h = \{(jh, 0)\} \cup \{(1, kh)\} \cup \{(jh, 1)\} \cup \{(0, kh)\}$$

$$\text{for } 0 \leq j, k \leq n+1.$$

Order the points of D_h in a left-to-right, down-to-up fashion. That is, if (j_1, k_1) and (j_2, k_2) are two gridpoints, then (j_1, k_1) precedes (j_2, k_2) if $k_1 < k_2$ or if $k_1 = k_2$ and $j_1 < j_2$.

Let u be an n^2 -vector whose components are ordered according to the order of D_h . Denote the components of u by $u_{j,k}$, (j, k) a gridpoint.

Let the matrix A defined by

$$\begin{aligned} (Au)_{j,k} = & B_{j,k} u_{j,k-1} + D_{j,k} u_{j-1,k} + E_{j,k} u_{j,k} \\ (2.2) \quad & + D_{j+1,k} u_{j+1,k} + B_{j,k+1} u_{j,k+1} \end{aligned}$$

be a discrete version of the differential operator defined in (2.1), and the equation

$$(Au)_{j,k} = q_{j,k}$$

be a discrete version of the boundary value problem in (2.1).

The important properties required of A are that

$$\begin{aligned}
 & \text{(i)} \quad B_{j,k}, D_{j,k} \leq 0; \\
 & \text{(ii)} \quad D_{j,k} = 0 \quad \text{for } j = 1, k = 1, 2, \dots, n; \\
 & \text{(iii)} \quad B_{j,k} = 0 \quad \text{for } k = 1, j = 1, 2, \dots, n; \\
 (2.3) \quad & \text{(iv)} \quad E_{j,k} = -(B_{j,k} + D_{j,k} + D_{j+1,k} + B_{j,k+1}) > 0, \\
 & \quad \text{if } D_{j,k} \neq 0 \quad \text{and} \quad B_{j,k} \neq 0; \\
 & \text{(v)} \quad E_{j,k} > -(B_{j,k} + D_{j,k} + D_{j+1,k} + B_{j,k+1}), \\
 & \quad \text{if } D_{j,k} = 0 \quad \text{or} \quad B_{j,k} = 0.
 \end{aligned}$$

These properties result from any of the common discretizations of the quantity

$$\frac{\partial}{\partial x_i} (a_i(x) \frac{\partial}{\partial x_i} u(x)).$$

For example,

$$\begin{aligned}
 \frac{\partial}{\partial x_i} (a_i(x) \frac{\partial}{\partial x_i} u(x)) & \approx h^{-2} \{ a_i(x + \frac{h}{2} e_i) [u(x) - u(x + h e_i)] \\
 & + a_i(x - \frac{h}{2} e_i) [u(x) - u(x - h e_i)] \},
 \end{aligned}$$

where e_1, e_2 are the unit vectors along the x_1 and x_2 axes.

This particular discretization yields,

$$\begin{aligned}
 B_{j,k} &= -a_2(jh, (k - \frac{1}{2})h), \\
 D_{j,k} &= -a_1((j - \frac{1}{2})h, kh), \\
 E_{j,k} &= a_2(jh, (k - \frac{1}{2})h) + a_2(jh, (k + \frac{1}{2})h) \\
 &\quad + a_1((j - \frac{1}{2})h, kh) + a_1((j + \frac{1}{2})h, kh).
 \end{aligned}
 \tag{2.4}$$

When $j = 1$ or $k = 1$, we adopt the convention that $B_{j,k} = 0$ or $D_{j,k} = 0$ respectively. Relations (2.3) follow.

3. THE FACTORIZATION

The difficulty in using Gaussian elimination to solve $Ax = b$ is that forward reduction transforms A with only five non-zero diagonals to an upper triangular matrix with n non-zero diagonals. Since Gaussian elimination is equivalent to factoring A as $A = L \cdot U$ where L is lower triangular and U is upper triangular with its main diagonal consisting of 1's, and since the factorization is unique, it follows that A cannot be factored as the product of sparse lower and upper triangular matrices.

The idea of a strongly implicit procedure is to replace the matrix A with a matrix of the form $A + B$, where $A + B$ can be written as the product of sparse lower and upper triangular matrices

$$A + B = L \cdot U.$$

The SI procedure of Stone constructs lower and upper triangular matrices, each with three non-zero diagonals in the same positions as the non-zero diagonals of A . That is,

$$(Lu)_{j,k} = b_{j,k} u_{j,k-1} + c_{j,k} u_{j-1,k} + d_{j,k} u_{j,k} \quad (3.1)$$

$$(Uu)_{j,k} = u_{j,k} + e_{j,k} u_{j+1,k} + f_{j,k} u_{j,k+1}$$

As a preliminary to the formal statement of the algorithm for computing L and U , observe that the product $L \cdot U$ is given by

$$\begin{aligned}
[(A + B)u]_{j,k} &= [(L \cdot U)u]_{j,k} = b_{j,k} u_{j,k-1} \\
&+ b_{j,k} \cdot e_{j,k-1} u_{j+1,k-1} + c_{j,k} u_{j-1,k} \\
(3.2) \quad &+ (d_{j,k} + b_{j,k} \cdot f_{j,k-1} + c_{j,k} e_{j-1,k}) u_{j,k} \\
&+ d_{j,k} \cdot e_{j,k} u_{j+1,k} + c_{j,k} \cdot f_{j-1,k} u_{j-1,k+1} \\
&+ d_{j,k} \cdot f_{j,k} u_{j,k+1}.
\end{aligned}$$

Let $0 \leq \alpha \leq 1$. Stone suggests in [4] the following algorithm for computing L and U :

$$\begin{aligned}
b_{j,k} &= B_{j,k} - \alpha c_{j,k-1} \cdot f_{j-1,k-1} \\
c_{j,k} &= D_{j,k} - \alpha b_{j-1,k} \cdot e_{j-1,k-1} \\
d_{j,k} + b_{j,k} \cdot f_{j,k-1} + c_{j,k} \cdot e_{j-1,k} &= E_{j,k} \\
(3.3) \quad &+ \alpha c_{j,k-1} \cdot f_{j-1,k-1} + \alpha b_{j-1,k} \cdot e_{j-1,k-1} \\
d_{j,k} \cdot e_{j,k} &= D_{j+1,k} - \alpha b_{j,k} \cdot e_{j,k-1} \\
d_{j,k} \cdot f_{j,k} &= B_{j,k+1} - \alpha c_{j,k} \cdot f_{j-1,k} \quad j,k = 1, 2, \dots, n.
\end{aligned}$$

Observe that the left side of (3.3) forms the elements of the (j,k) row of $L \cdot U$.

Stone proved in [4] that $L \cdot U$ is symmetric and his proof is reproduced here. First, observe that in the actual computational algorithm it is not necessary to compute $b_{j,k}$ and $c_{j,k}$ since it can be seen by adjusting subscripts in (3.3) that

$$(3.4) \quad b_{j,k} = d_{j,k-1} \cdot f_{j,k-1}$$

and

$$(3.5) \quad c_{j,k} = d_{j-1,k} \cdot e_{j-1,k}.$$

Changing the subscripts in (3.5) gives

$$(3.6) \quad c_{j+1,k-1} = d_{j,k-1} \cdot e_{j,k-1}.$$

Multiplying the right hand side of (3.6) by the left side of (3.4) gives

$$b_{j,k}(d_{j,k-1} \cdot e_{j,k-1}) = c_{j+1,k-1}(d_{j,k-1} \cdot f_{j,k-1})$$

or

$$(3.7) \quad b_{j,k} \cdot e_{j,k-1} = c_{j+1,k-1} \cdot f_{j,k-1}$$

Equation (3.7), along with (3.4) and (3.5), establishes the symmetry of $A + B$ as defined in (3.2).

We now prove a lemma which establishes a useful set of relations between the elements of A and the elements of $A + B$.

Lemma 1:

The quantities defined in equation (3.3) exist and satisfy the following relations:

$$(a) \quad b_{j,k} \leq B_{j,k} \leq 0$$

$$(b) \quad c_{j,k} \leq D_{j,k} \leq 0$$

$$(c) \quad d_{j,k} > 0$$

(3.8)

$$(d) \quad -1 \leq e_{j,k} \leq 0$$

$$(e) \quad -1 \leq f_{j,k} \leq 0$$

$$(f) \quad e_{j,k} + f_{j,k} + 1 \geq 0$$

Proof:

The proof is by induction. For $j = k = 1$, we have,

$$b_{1,1} = B_{1,1} = 0,$$

$$c_{1,1} = D_{1,1} = 0,$$

$$d_{1,1} = E_{1,1} > 0,$$

$$-1 \leq e_{1,1} = \frac{D_{2,1}}{E_{1,1}} \leq 0,$$

$$-1 \leq f_{1,1} = \frac{B_{1,2}}{E_{1,1}} \leq 0,$$

and

$$e_{1,1} + f_{1,1} + 1 = \frac{D_{2,1}}{E_{1,1}} + \frac{B_{1,2}}{E_{1,1}} + 1 = \frac{D_{2,1} + B_{1,2} + E_{1,1}}{E_{1,1}} > 0.$$

Now assume that the relations (a) - (f) hold for all points preceding (j,k) .

To prove (3.8) (a) and (b), we have

$$c_{j,k-1} \leq 0, \quad f_{j-1,k-1} \leq 0, \quad b_{j-1,k} \leq 0,$$

$$e_{j-1,k-1} \leq 0 \quad \text{and} \quad 0 \leq \alpha \leq 1.$$

It follows that

$$b_{j,k} = B_{j,k} - \alpha c_{j,k-1} \cdot f_{j-1,k-1} \leq B_{j,k} \leq 0$$

and

$$c_{j,k} = D_{j,k} - \alpha b_{j-1,k} \cdot e_{j-1,k-1} \leq D_{j,k} \leq 0.$$

To prove (3.8) (c) we have, by assumption, $1 + f_{j,k-1} \geq 0$ and $1 + e_{j-1,k} \geq 0$. Therefore,

$$\begin{aligned} d_{j,k} &= E_{j,k} - b_{j,k} \cdot f_{j,k-1} - c_{j,k} \cdot e_{j-1,k} \\ &\quad + \alpha c_{j,k-1} \cdot f_{j-1,k-1} + \alpha b_{j-1,k} \cdot e_{j-1,k-1} \\ &= E_{j,k} + B_{j,k} + D_{j,k} - b_{j,k}(1 + f_{j,k-1}) \\ &\quad - c_{j,k}(1 + e_{j-1,k}) > 0. \end{aligned}$$

Next we prove (3.8) (d) and (e).

Certainly,

$$e_{j,k} = \frac{D_{j+1,k} - \alpha b_{j,k} e_{j,k-1}}{d_{j,k}} \leq 0$$

and

$$f_{j,k} = \frac{B_{j,k+1} - \alpha c_{j,k} \cdot f_{j-1,k}}{d_{j,k}} \leq 0.$$

To prove that -1 is a lower bound for these quantities, observe that

$$e_{j,k+1} = \frac{e_{j,k} \cdot d_{j,k} + d_{j,k}}{d_{j,k}}.$$

Since (3.8) (d) and (e) are valid for $(j,k-1)$, it follows from (3.3) that

$$\begin{aligned} e_{j,k} \cdot d_{j,k} + d_{j,k} &= D_{j+1,k} - \alpha b_{j,k} \cdot e_{j,k-1} + d_{j,k} \\ &= E_{j,k} + D_{j+1,k} - \alpha b_{j,k} \cdot e_{j,k-1} - b_{j,k} \cdot f_{j,k-1} \\ &\quad - c_{j,k} \cdot e_{j-1,k} + \alpha c_{j,k-1} \cdot f_{j-1,k-1} + \alpha b_{j-1,k} \cdot e_{j-1,k-1} \\ &= E_{j,k} + D_{j+1,k} + B_{j,k} + D_{j,k} - b_{j,k} (1 + \alpha e_{j,k-1} + f_{j,k-1}) \\ &\quad - c_{j,k} (1 + e_{j-1,k}) \geq 0. \end{aligned}$$

Therefore, $-1 \leq e_{j,k}$.

Similarly, $-1 \leq f_{j,k} \leq 0$.

To prove the remaining assertion, we have from (3.3)

$$\begin{aligned} e_{j,k} + f_{j,k} + 1 &= \frac{D_{j+1,k} - \alpha b_{j,k} \cdot e_{j,k-1}}{d_{j,k}} \\ &\quad + \frac{B_{j,k+1} - \alpha c_{j,k} \cdot f_{j-1,k}}{d_{j,k}} + 1 \end{aligned}$$

$$\begin{aligned}
&= \frac{d_{j,k} + D_{j+1,k} + B_{j,k+1} - \alpha b_{j,k} \cdot e_{j,k-1} - \alpha c_{j,k} \cdot f_{j-1,k}}{d_{j,k}} \\
&= \frac{E_{j,k} + D_{j+1,k} + B_{j,k+1} - \alpha b_{j,k} \cdot e_{j,k-1} - \alpha c_{j,k} \cdot f_{j-1,k} - b_{j,k} \cdot f_{j,k-1}}{d_{j,k}} \\
&+ \frac{-c_{j,k} \cdot e_{j-1,k} + \alpha c_{j,k-1} \cdot f_{j-1,k-1} + \alpha b_{j-1,k} \cdot e_{j-1,k-1}}{d_{j,k}} \\
&= \frac{E_{j,k} + B_{j,k} + D_{j,k} + B_{j,k+1} + D_{j+1,k}}{d_{j,k}} - \frac{b_{j,k}}{d_{j,k}} (1 + \alpha e_{j,k-1} + f_{j,k-1}) \\
&- \frac{c_{j,k}}{d_{j,k}} (1 + \alpha f_{j-1,k} + e_{j-1,k})
\end{aligned}$$

By the induction hypothesis, and from the facts that

$$1 + \alpha e_{j,k-1} + f_{j,k-1} \geq 1 + e_{j,k-1} + f_{j,k-1} \geq 0$$

and

$$1 + \alpha f_{j-1,k} + e_{j-1,k} \geq 1 + e_{j-1,k} + f_{j-1,k} \geq 0,$$

it follows that

$$e_{j,k} + f_{j,k} + 1 \geq 0.$$

4. ANALYSIS AND CONVERGENCE CONDITIONS

The iteration we consider is defined as follows:

$$(4.1) \quad (A + B)u_{n+1} = (A + B)u_n - \tau(Au_n - q).$$

Our objectives in this section are to prove that $A + B$ is positive definite and that (4.1) converges for $\tau \in (0, 2/\|A\|_2)$. First, we establish the notation and conventions that we will use. Let n denote the number of grid points on a horizontal or vertical line. Let $N = n^2$. In the remainder of this thesis, it is convenient to rely on standard matrix notation. Thus, if $C = (c_{ij})$ is a matrix, $1 \leq i, j \leq n$, then c_{ij} is the element at the i -th row and the j -th column of the matrix, and the notation c_{ij} no longer refers to an element of the row of the matrix corresponding to the gridpoint defined by the intersection of the i -th horizontal line with the j -th vertical line.

The next lemma provides a convenient representation of the quadratic forms $\langle Au, u \rangle$ and $\langle Bu, u \rangle$.

Lemma 2:

For the matrix A , whose properties are listed in (2.3), and for B defined by $B = LU - A$, we have

$$\begin{aligned} \langle Au, u \rangle = & \sum_{i=1}^{N-1} -a_{i,i+1} |u_{i+1} - u_i|^2 + \sum_{i=1}^{N-n} -a_{i,i+n} |u_{i+n} - u_i|^2 \\ & + \sum_{i=1}^N (a_{i,i} + a_{i,i+1} + a_{i,i+n} + a_{i-1,i} + a_{i-n,i}) |u_i|^2 \end{aligned}$$

and

$$\begin{aligned}
\langle Bu, u \rangle &= \alpha \sum_{i=2}^{N-n} b_{i,i+n} |u_{i+n} - u_i|^2 - \alpha \sum_{i=2}^{N-n} b_{i,i+n} |u_i|^2 \\
&+ \alpha \sum_{i=n+1}^{N-1} b_{i,i+1} |u_{i+1} - u_i|^2 - \alpha \sum_{i=n+1}^{N-1} b_{i,i+1} |u_i|^2 \\
&- \sum_{i=1}^{N-n-1} b_{i+1,i+n} |u_{i+1} - u_{i+n}|^2 \\
&+ \sum_{i=1}^{N-n-1} b_{i+1,i+n} (|u_{i+1}|^2 + |u_{i+n}|^2).
\end{aligned}$$

Proof:

We only prove the second part of the assertion. The first part is a familiar fact.

We have

$$\begin{aligned}
\langle Bu, u \rangle &= -\alpha \sum_{i=n+2}^N b_{i,i-n} \cdot u_{i-n} \cdot \bar{u}_i \\
&+ \sum_{i=n+1}^{N-1} b_{i,i+1-n} \cdot u_{i+1-n} \cdot \bar{u}_i \\
&- \alpha \sum_{i=n+2}^N b_{i,i-1} \cdot u_{i-1} \cdot \bar{u}_i \\
&+ \alpha \sum_{i=n+2}^N b_{i,i} |u_i|^2 \\
&- \alpha \sum_{i=n+1}^{N-1} b_{i,i+1} \cdot u_{i+1} \cdot \bar{u}_i
\end{aligned}$$

$$+ \sum_{i=2}^{N-n} b_{i,i+n-1} \cdot u_{i+n-1} \cdot \bar{u}_i$$

$$- \alpha \sum_{i=2}^{N-n} b_{i,i+n} \cdot u_{i+n} \cdot \bar{u}_i.$$

Since B is symmetric, and since

$$b_{i,i} = b_{i,i-n} + b_{i,i-1},$$

$$\langle Bu, u \rangle = - \alpha \sum_{i=n+2}^N b_{i-n,i} \cdot u_{i-n} \cdot \bar{u}_i$$

$$+ \sum_{i=n+1}^{N-1} b_{i+1-n,i} \cdot u_{i+1-n} \cdot \bar{u}_i$$

$$- \alpha \sum_{i=n+2}^N b_{i-1,i} \cdot u_{i-1} \cdot \bar{u}_i$$

$$+ \alpha \sum_{i=n+2}^N (b_{i-n,i} + b_{i-1,i}) |u_i|^2$$

$$- \alpha \sum_{i=n+1}^{N-1} b_{i,i+1} \cdot u_{i+1} \cdot \bar{u}_i$$

$$+ \sum_{i=2}^{N-n} b_{i,i+n-1} \cdot u_{i+n-1} \cdot \bar{u}_i$$

$$- \alpha \sum_{i=2}^{N-n} b_{i,i+n} \cdot u_{i+n} \cdot \bar{u}_i$$

In the expression on the right, change the variable of summation in the first sum and combine with the last sum. Also, change the variable of summation in the third sum and combine with the fifth sum. Finally, change the variable of summation in the second and the sixth sums.

We have,

$$\begin{aligned}
 \langle Bu, u \rangle = & -\alpha \sum_{i=2}^{N-n} b_{i,i+n} (u_i \cdot \bar{u}_{i+n} + u_{i+n} \cdot \bar{u}_i) \\
 & -\alpha \sum_{i=n+1}^{N-1} b_{i,i+1} (u_i \cdot \bar{u}_{i+1} + u_{i+1} \cdot \bar{u}_i) \\
 & + \sum_{i=1}^{N-n-1} b_{i+1,i+n} (u_{i+1} \cdot \bar{u}_{i+n} + u_{i+n} \cdot \bar{u}_{i+1}) \\
 & + \alpha \sum_{i=n+2}^N (b_{i-n,i} + b_{i-1,i}) |u_i|^2.
 \end{aligned}$$

From the identities

$$-u_{i+n} \bar{u}_i - u_i \bar{u}_{i+n} = |u_{i+n} - u_i|^2 - (|u_{i+n}|^2 + |u_i|^2),$$

$$-u_{i+1} \bar{u}_i - u_i \bar{u}_{i+1} = |u_{i+1} - u_i|^2 - (|u_{i+1}|^2 + |u_i|^2)$$

and

$$u_{i+1} \bar{u}_{i+n} + u_{i+n} \bar{u}_{i+1} = -|u_{i+1} - u_{i+n}|^2 + |u_{i+1}|^2 + |u_{i+n}|^2,$$

we have

$$\begin{aligned}
\langle Bu, u \rangle &= \alpha \sum_{i=2}^{N-n} b_{i,i+n} |u_{i+n} - u_i|^2 \\
&\quad - \alpha \sum_{i=2}^{N-n} b_{i,i+n} (|u_{i+n}|^2 + |u_i|^2) \\
&\quad + \alpha \sum_{i=n+1}^{N-1} b_{i,i+1} |u_{i+1} - u_i|^2 \\
&\quad - \alpha \sum_{i=n+1}^{N-1} b_{i,i+1} (|u_{i+1}|^2 + |u_i|^2) \\
&\quad - \sum_{i=1}^{N-n-1} b_{i+1,i+n} |u_{i+1} - u_{i+n}|^2 \\
&\quad + \sum_{i=1}^{N-n-1} b_{i+1,i+n} (|u_{i+1}|^2 + |u_{i+n}|^2) \\
&\quad + \alpha \sum_{i=2}^{N-n} b_{i,i+n} |u_{i+n}|^2 \\
&\quad + \alpha \sum_{i=n+1}^{N-1} b_{i,i+1} |u_{i+1}|^2.
\end{aligned}$$

It follows that

$$\begin{aligned}
\langle Bu, u \rangle &= \alpha \sum_{i=2}^{N-n} b_{i,i+n} |u_{i+n} - u_i|^2 \\
&\quad - \alpha \sum_{i=2}^{N-n} b_{i,i+n} |u_i|^2
\end{aligned}$$

$$+ \alpha \sum_{i=n+1}^{N-1} b_{i,i+1} |u_{i+1} - u_i|^2$$

$$- \alpha \sum_{i=n+1}^{N-1} b_{i,i+1} |u_i|^2$$

$$- \sum_{i=1}^{N-n-1} b_{i+1,i+n} |u_{i+1} - u_{i+n}|^2$$

$$+ \sum_{i=1}^{N-n-1} b_{i+1,i+n} (|u_{i+1}|^2 + |u_{i+n}|^2),$$

which was to be proved.

The following lemma of Dupont, Kendall and Rachford [1,2] will be used in the proof of theorem 1.

Lemma 3:

Let c_i be nonnegative for $i = 1, \dots, n$. Let $\sum_{i=1}^n c_i$ be positive.

Let e and a_i , $i = 1, 2, \dots, n$ be complex.

Then

$$\left[\sum_{i=1}^n c_i \right]^{-1} \sum_{1 \leq i \leq \ell \leq n} c_i c_\ell |a_i - a_\ell|^2 \leq \sum_{i=1}^n c_i |a_i - e|^2$$

Theorem 1:

For $0 \leq \alpha \leq 1$ the matrix $A+B$, defined by (3.2), is positive definite.

Proof:

Let $A = (a_{ij})$.

We have from lemma 2,

$$\begin{aligned}
\langle (A+B)u, u \rangle &= \sum_{i=1}^{N-1} -a_{i,i+1} |u_{i+1} - u_i|^2 + \sum_{i=1}^{N-n} -a_{i,i+n} |u_{i+n} - u_i|^2 \\
&+ \sum_{i=1}^N (a_{i,i} + a_{i,i+1} + a_{i,i+n} + a_{i-1,i} + a_{i-n,i}) |u_i|^2 \\
&+ \alpha \sum_{i=n+1}^{N-1} b_{i,i+1} |u_{i+1} - u_i|^2 - \alpha \sum_{i=n+1}^{N-1} b_{i,i+1} |u_i|^2 \\
&+ \alpha \sum_{i=2}^{N-n} b_{i,i+n} |u_{i+n} - u_i|^2 - \alpha \sum_{i=2}^{N-n} b_{i,i+n} |u_i|^2 \\
&+ \sum_{i=1}^{N-n-1} b_{i+1,i+n} (|u_{i+1}|^2 + |u_{i+n}|^2) \\
&- \sum_{i=1}^{N-n-1} b_{i+1,i+n} |u_{i+1} - u_{i+n}|^2.
\end{aligned}$$

Since $b_{i,i+1} = 0$ for $i = 1, 2, \dots, n$, and $b_{1,1+n} = 0$,

$$\begin{aligned}
\langle (A+B)u, u \rangle &= \sum_{i=1}^{N-1} (\alpha b_{i,i+1} - a_{i,i+1}) |u_{i+1} - u_i|^2 \\
&+ \sum_{i=1}^{N-n} (\alpha b_{i,i+n} - a_{i,i+n}) |u_{i+n} - u_i|^2 \\
&+ \sum_{i=1}^N (a_{i,i} + a_{i,i+1} + a_{i,i+n} + a_{i-1,i} + a_{i-n,i}) |u_i|^2
\end{aligned}$$

$$\begin{aligned}
& - \sum_{i=1}^{N-n-1} b_{i+1,i+n} |u_{i+1} - u_{i+n}|^2 \\
& + \sum_{i=2}^{N-n} (b_{i,i+n-1} - \alpha b_{i,i+n}) |u_i|^2 \\
& + \sum_{i=n+1}^{N-1} (b_{i,i+1-n} - \alpha b_{i,i+1}) |u_i|^2.
\end{aligned}$$

By lemma 3, applied to the first two sums, and since $b_{N-n+1,N} = 0$,

$$\begin{aligned}
\langle (A+B)u, u \rangle & \geq \sum_{i=1}^{N-n} \left\{ \frac{(\alpha b_{i,i+1} - a_{i,i+1})(\alpha b_{i,i+n} - a_{i,i+n})}{\alpha b_{i,i+1} - a_{i,i+1} + \alpha b_{i,i+n} - a_{i,i+n}} - b_{i+1,i+n} \right\} \\
& \cdot |u_{i+1} - u_{i+n}|^2 + \sum_{i=N-n+1}^{N-1} (\alpha b_{i,i+1} - a_{i,i+1}) |u_{i+1} - u_i|^2 \\
(4.2) \quad & + \sum_{i=1}^N (a_{i,i} + a_{i,i+1} + a_{i,i+n} + a_{i-1,i} + a_{i-n,i}) |u_i|^2 \\
& + \sum_{i=2}^{N-n} (b_{i,i+n-1} - \alpha b_{i,i+n}) |u_i|^2 \\
& + \sum_{i=n+1}^{N-1} (b_{i,i+1-n} - \alpha b_{i,i+1}) |u_i|^2.
\end{aligned}$$

To prove that $\langle (A+B)u, u \rangle \geq 0$, consider each sum in the last expression separately. The coefficient of term i in the first sum is

$$\Gamma_i = \frac{(\alpha b_{i,i+1} - a_{i,i+1})(\alpha b_{i,i+n} - a_{i,i+n})}{\alpha b_{i,i+1} - a_{i,i+1} + \alpha b_{i,i+n} - a_{i,i+n}} - b_{i+1,i+n}.$$

It is necessary to replace the quantities in Γ_i by the expressions which define them. To do so, let matrix row i correspond to point (j,k) of the grid. Then it follows that

$$\Gamma_i = \frac{(\alpha b_{j,k} \cdot e_{j,k-1}^{-D_{j+1,k}})(\alpha c_{j,k} \cdot f_{j-1,k}^{-B_{j,k+1}})}{\alpha b_{j,k} \cdot e_{j,k-1}^{-D_{j+1,k}} + \alpha c_{j,k} \cdot f_{j-1,k}^{-B_{j,k+1}}} - d_{j,k} \cdot e_{j,k} \cdot f_{j,k}.$$

By (3.3),

$$\begin{aligned} (4.3) \quad \Gamma_i &= \frac{(-d_{j,k} \cdot e_{j,k})(-d_{j,k} \cdot f_{j,k})}{-d_{j,k} \cdot e_{j,k} - d_{j,k} \cdot f_{j,k}} - d_{j,k} \cdot e_{j,k} \cdot f_{j,k} = \\ &= \frac{d_{j,k} e_{j,k} f_{j,k} (1 + e_{j,k} + f_{j,k})}{-e_{j,k} - f_{j,k}}. \end{aligned}$$

Therefore, by lemma 1, $\Gamma_i \geq 0$.

In order to see that the arithmetic quantities and operations in the above reduction leading to (4.3) are defined, we have for denominator of Γ_i ,

$$\begin{aligned} \alpha b_{j,k} \cdot e_{j,k-1}^{-D_{j+1,k}} + \alpha c_{j,k} \cdot f_{j-1,k}^{-B_{j,k+1}} &\geq -D_{j+1,k} - B_{j,k+1}. \end{aligned}$$

Equality implies that the sum on the right is positive.

For the second sum on the right of (4.2) we have from (3.3),

$$\alpha b_{l,i+1} - a_{l,i+1} = \alpha b_{j,k} \cdot e_{j,k-1}^{-D_{j+1,k}} \geq 0,$$

where row l of the matrix corresponding to the (j,k) gridpoint.

The third sum in (4.2) is obviously nonnegative.

Consider the fourth and fifth sums in (4.2). Let row i correspond to the (j,k) gridpoint, as before. We have for the coefficients of these sums,

$$\begin{aligned} b_{i,i+n-1} - \alpha b_{i,i+n} &\equiv c_{j,k} \cdot f_{j-1,k} - \alpha c_{j,k} \cdot f_{j-1,k} \\ &= c_{j,k} \cdot f_{j-1,k} (1 - \alpha) \geq 0 \end{aligned}$$

and

$$\begin{aligned} b_{i,i+1-n} - \alpha b_{i,i+1} &\equiv b_{j,k} \cdot e_{j,k-1} - \alpha b_{j,k} \cdot e_{j,k-1} \\ &= b_{j,k} \cdot e_{j,k-1} (1 - \alpha) \geq 0. \end{aligned}$$

Therefore the fourth and fifth sums are nonnegative.

It follows that

$$\langle (A+B)u, u \rangle \geq 0.$$

The inequality is strict, since $A+B$ is nonsingular.

For,

$$\det(A+B) = \det(L) = \prod d_{j,k} \neq 0.$$

This completes the proof of Theorem 1.

We can now show there exist values of the parameter τ which assure convergence of the iteration scheme,

$$(A+B)u_{n+1} = (A+B)u_n - \tau(Au_n - q).$$

Let u satisfy

$$(A+B)u = (A+B)u - \tau(Au - q).$$

We have

$$(A+B)e_{n+1} = (A+B-\tau A)e_n$$

where $e_n = u - u_n$.

Let w_1, w_2, \dots, w_n be a complete set of eigenvectors of $(A+B)^{-1}(A+B-\tau A)$, orthonormal with respect to the inner product $(v_1, v_2)_{(A+B)}$ defined by $(v_1, v_2)_{(A+B)} = ((A+B)v_1, v_2)$. Let $\lambda_1, \lambda_2, \dots, \lambda_n$ be the corresponding eigenvalues.

Let

$$e_n = \sum_{i=1}^N d_i^{(n)} w_i$$

and

$$e_{n+1} = \sum_{i=1}^N d_i^{(n+1)} w_i.$$

Since

$$(A+B-\tau A)w_i = \lambda_i (A+B)w_i,$$

it follows that

$$\lambda_i = \lambda_i((A+B)w_i, w_i) = ((A+B-\tau A)w_i, w_i) = 1 - \tau(Aw_i, w_i).$$

Let M and m be the largest and smallest eigenvalues of A respectively,

so that

$$1 - M \leq \lambda_i = 1 - \tau(Aw_i, w_i) \leq 1 - \tau m.$$

Therefore,

$$|\lambda_i| = |1 - \tau(Aw_i, w_i)| < 1$$

if $1 - \tau M > -1$ and $1 - \tau m < 1$, i.e., if $0 < \tau < \frac{2}{M}$.

For this range of τ ,

$$\|u - u_{n+1}\|_{A+B} < q \|u - u_n\|_{A+B}$$

where

$$q = \min(|1 - \tau M|, |1 - \tau m|),$$

and convergence results.

We have proved

Theorem 2:

Let A be the positive definite real symmetric matrix defined in (2.2), with largest and smallest eigenvalues M and m respectively. Then the iteration (4.1) converges if the iteration parameter τ satisfies

$$0 < \tau < \frac{2}{M}$$

and the $(A+B)$ - norm of the error is reduced by at least the quantity

$$q = \min(|1 - \tau M|, |1 - \tau m|)$$

at each step.

Dupont, Rachford and Kendall proved a lemma in [2] which also gives bounds on values of τ which imply convergence. Their bounds depend on the auxiliary matrix B , whereas of course the bounds 0 and $2/\|A\|_2$, above, do not.

5. COMPUTATIONAL NOTES

We give below a version of the iteration

$$(A+B)u_{n+1} = (A+B)u_n - \tau(Au_n - q)$$

suitable for coding, and give a count of the number of operations required for each step.

Stone suggests, [5], letting

$$R_n = q - Au_n$$

and

$$\delta_{n+1} = u_{n+1} - u_n.$$

Solve

$$L t = R_n$$

and

$$U \delta_{n+1} = t.$$

Set

$$u_{n+1} = u_n + \delta_{n+1}.$$

Note that for $\tau = 1$, if the initial approximation, x_0 , is $x_0 = 0$, then x_1 is the solution of $(A+B)x_1 = q$. It is plausible that x_1 is a good approximation to the solution of $Ax = q$.

If A is of order N , the factorization of A into the product of L and U requires $O(13N)$ operations.

For the iterative procedure defined above:

- (a) To solve $L t = R$ requires $O(3N)$ operations;
- (b) To solve $U \delta = t$ requires $O(2N)$ operations;
- (c) To determine R_n requires $O(5N)$ operations.

Therefore there are $O(10N)$ operations for each iteration.

LIST OF REFERENCES

- [1] T. Dupont, "A Factorization Procedure for the Solution of Elliptic Difference Equations," SIAM Journal on Numerical Analysis, December, 1968.
- [2] T. Dupont, R. P. Kendall, and H. H. Rachford, Jr., "An Approximate Factorization Procedure for Solving Self-adjoint Elliptic Difference Equations," SIAM Journal On Numerical Analysis, September, 1968.
- [3] E. G. D'Yakonov, "An Iteration Method for Solving Systems of Finite Difference Equations," Dokl. Akad. Nauk. SSSR 138, 1961.
- [4] H. L. Stone, Private communication, April, 1969.
- [5] H. L. Stone, "Iterative Solution of Implicit Approximations of Multidimensional Partial Differential Equations," SIAM Journal On Numerical Analysis, September, 1968.
- [6] R. S. Varga, Matrix Iterative Analysis, Prentice-Hall, Englewood Cliffs, New Jersey, 1962.
- [7] E. L. Wachspress, Iterative Solution of Elliptic Systems, Prentice-Hall, Englewood Cliffs, New Jersey, 1966.

NOV 22 1972

MAY 10 1973



UNIVERSITY OF ILLINOIS-URBANA
510.84 IL6R no. C002 no. 391-396(1970)
Digital computer internal report /



3 0112 088399149